M.TECH/BT/2^{ND} SEM/BIOT 5201/2023

ADVANCED BIOINFORMATICS
(BIOT 5201)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and
any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) In the SCOP database of 2020 what were the main improvements made over the earlier version?
(a) Increases coverage of protein structural data
(b) Improves the schema of database coverage
(c) Includes genomic information
(d) Both (a) and (b).

(ii) The problem with the ab-initio prediction method for protein structure prediction is represented by which one of the following choices?
(a) It is able to predict protein tertiary structures from first principles, in the absence of any structural information
(b) It is able to predict protein tertiary structures but has to predict secondary structure first
(c) Even short polypeptide chains can fold into a potentially infinite number of different structures
(d) All of the above.

(iii) Phylogenetic relationship can be shown by
(a) Dendrogram (b) GENBANK
(c) data retrieving tool (d) data search tool.

(iv) In protein secondary structure prediction artificial neural networks are used. Which of the following input-output combinations are favoured in such a case?
(a) Step function input-step function output
(b) Step function input-sigmoidal function output
(c) Bi-directional input
(d) Both (b) and (c).

(v) The process of finding the relative location of gene on a chromosome is referred to by which one of the following terms?
(a) Gene tracing (b) Gene mapping
(c) Genome walking (d) Chromosome walking.

(vi) Once complete energy minimization is completed for a small molecule binding to a receptor target, which one of the following options are valid?
(a) Ligand, receptor are in their lowest energy, most stable conformations
(b) Ligand and ligand-receptor complex are in their most stable conformations
(c) The most stable, lowest energy ligand-receptor complex may not be the bioactive one
(d) All of the above.

(vii) To use molecular data to reconstruct evolutionary history requires making a number of reasonable assumptions. Which of the following is incorrect about it?
(a) The molecular sequences used in phylogenetic construction are homologous
(b) The molecular sequences used in phylogenetic construction share a common origin
(c) Phylogenetic divergence cannot be bifurcating
(d) Parent branch splits into two daughter branches at any given point.

(viii) A limited energy minimization*is sometimes required* for which of the following structure building modalities?
(a) Homology modelling using structure-template
(b) Threading using homology between native folds
(c) Ab-initio using short templates
(d) Ab-initio from protein folding principles.

(ix) Which of the following is not a variant of BLAST?

(a) BLASTN (b) BLASTP

(c) BLASTX (d) TBLASTNX.

(x) The Tanimoto coefficient T is frequently used as an index to quantify structural similarity between potential therapeutic drug structures; T is defined by which of the following expressions?

(a) $T=N_{\sigma}-N_{\beta}$ (b) $T= N_{11}/n-N_{00}$

(c) T=clog P-524 (d) T= S ln P.

Group - B

2. (a) What is the basis of *ab initio* based gene prediction? What are the two important conceptual steps in ab-initio gene prediction? *(CO4)(Understand-LOCQ)*

(b) Itemize the involvement of different gene signals in this procedure. *(CO4)(Understand-analyze/IOCQ)*

(c) "Presence of just a start codon is sufficient to initiate the beginning of the frame of translation". Using an example from bacterial gene prediction, evaluate this statement on a scientific-technical basis. *(CO4)(Evaluate-HOCQ)*

(d) 'In order to evaluate the accuracy of a prediction program (for genes, proteins), a performance evaluation of the said program is a requirement. What are the two parameters that are essential for this performance evaluation? Define the parameters mathematically briefly explaining their significance. How does one quantitatively summarize the two parameters into a single "summarizing parameter"? Explain your answer. *(CO4)(Understand-apply -IOCQ)*

3 + 2 + 3 + 4 = 12

3. (a) A protein sequence has been provided to you. The stated purpose is to make an attempt at structural and functional analysis of this newly determined sequence by sequence comparison.

(i) What are the type(s) of sequence alignment that you will adopt? Outline the alignment procedure stepwise with the help of a diagram. *[(CO3)(Remember-Understand)/LOCQ)]*

(b) The alignment results from part (a) are analyzed using E value and P values. What are the ranges of these values and discuss their inter-relationship. *[(CO3)(Understand-analyze/IOCQ)]*

6 + 6 = 12

Group - C

4. (a) A, B, C, D are four taxa whose distances are given: AB=0.40, AC=0.35, AD=0.60, BC=0.45, BD=0.70 and CD=0.55 based on suitable alignment method construct the phylogenetic tree. Show step wise how the final tree is developed. *(CO3)(Construct-HOCQ)*

(b) Name one bioinformatics software tool that is based on the clustering method you adopted. Enumerate the advantages and disadvantages of the method. *(CO3)(analyze IOCQ)*

6 + (1 + 2 + 3) = 12

5. (a) What is the dual basis of a PSI-BLAST program? Use a schematic flowchart of a PSI-BLAST calculation to detect protein sequences in a database similar to a probe sequence. *(CO3)(Understand-apply –IOCQ)*

(b) Use the steps in the schematic profile to discuss how you would overcome profile drift in a PSI-DRIFT calculation. *(CO3)(Evaluate –IOCQ)*

[(CO3)(Analyze/HOCQ)]

(2 + 5) + 5 = 12

Group - D

6. (a) Why are protein secondary structure prediction algorithms considered more accurate when combined with multiple sequence alignment (MSA) and neural networks (NN)? Limit your answer with two specific points. How much is the accuracy percentage improved? Use a flowchart explaining the steps to represent the operation of a neural network that is applied only to the secondary structure prediction of a protein. What filtering effect (in structural terms) happens when the transition occurs from the last hidden layer of a neural network to the final jury layer in a protein secondary structure algorithm? *[(CO4)(Understand-analyze/IOCQ)]*

(b) Transcriptomics deals with the inventory of RNA molecules in the cell. Why is detailed structural information about RNA important? Outline the reasons that make tertiary structure prediction of RNA difficult. Draw a schematic diagram of a hypothetical RNA molecule? What are the two methods of RNA secondary structure prediction and briefly describe the principles behind them. Itemize the steps in the ab initio approach of RNA structure prediction. *[(CO4)(Understand-explain/IOCQ]*

(4 + 2) + (4 + 2) = 12

7. (a) Represent in a schematic form the steps in an artificial neural network as a machine learning process using a globular protein sequence as an input, the purpose being to obtain a 3 state prediction of the secondary structure of a protein. Comment specifically on the improvement of prediction accuracy when a multiple sequence alignment (MSA) is combined with a neural network. *[(CO4)(Understand-analyze/IOCQ)]*

(b) Use two diagrams of RNA secondary structure to show all special loops and other super-secondary structures are present. Use a table to show the differences between the two main algorithms that are used for RNA secondary structure prediction. How is correlation coefficient calculated for such an algorithm's prediction? *[(CO4)(Understand-analyze/IOCQ)]*

(c) Using the SCOP database as an example, represent the characteristics of a database. In using Linnaean taxonomy, how is SCOP 's structure different from other protein classification databases? Briefly explain the algorithm that forms the basis of the comparable CATH protein classification database. *[(CO2)(Understand-apply/IOCQ)]*

4 + 4 + 4 = 12

Group - E

8. (a) "Methods of bioinformatics assist in selecting targets for structure determination by experimental (wet lab based) based methods." What is the basis of such selection? Name two such wet lab based methods and two methods based in bioinformatics". Itemize the scientific -technical goals of "target selection". Choose one example to explain any one of the cited goals of protein target selection. *[(CO5)(Understand-analyze/IOCQ)]*

(b) Itemize the reasons why the process of energy minimization plays a more important role in a macromolecule like a protein. Use a labelled potential energy diagram for a macromolecule to explain your answer. *[(CO3)(Understand-explain/IOCQ)]*

(c) Many proteins have been isolated from pathogens that have corresponding human homologues. Method development developed in human biology led to new method for comparison of parameters for specificity determination in the binding sites of two homologous proteins. How would you utilize this method for finding new targets? *[(CO4)(Analyse/HOCQ)]*

5 + 4 + 3 = 12

9. (a) "The combinatorial chemistry and molecular diversity (CCMD) based technologies have found use in a wide variety of drug discovery and development related applications." Use a flow chart to represent the critical technologies in the combinatorial chemistry process. In what segment of the library design flowchart does "bioinformatics" *broadly defined* plays a critically important role? Explain your answer. Use a *reaction diagram* to explain how three Building Block (BB) families can give more than 1.5 million compounds in 2 steps. Show your calculations. *[CO6(Understand-apply/IOCQ]*

(b) Consider a small molecule ligand, L, that is being developed as a lead compound for therapeutic purposes. The particular protein, P, involved in the disease progression has been identified and its sequence and structure are known. How would you computationally define and represent this binding process? Itemize both the physico-chemical parameters and map these against the computer based formalisms. Your answer should incorporate the different ways you can computationally represent the chemical possibilities in the P+L binary binding process. *[(CO6)(Understand-apply-analyse/IOCQ)]*

6 + 6 = 12

Cognition Level	*LOCQ*	*IOCQ*	*HOCQ*
Percentage distribution	*9.38*	*72.92*	*17.70*

Course Outcome (CO):

After the completion of the course students will be able to

- use acquired knowledge about different bioinformatics experiment categories (e.g. sequence, structure analysis) and their applications in new biology (e.g. genomics, proteomics)
- learn organization and characterization of primary and specialized databases and portals, introduction to new applications of databases/portals , introductions to new applications of databases/portals towards study of metabolic pathways and systems biology
- learn and apply sequence alignment methodologies (including comparison of applicable heuristic and dynamic algorithms) for pairwise and multiple sequence alignment and molecular phylogenetics
- learn and apply bioinformatics based software tools (and the algorithms underlying them) for annotation and structure prediction of prokaryotic and eukaryotic genes, RNA secondary structure prediction and secondary structure prediction of globular, fibrous and membrane proteins (e.g. use of artificial neural networkand Hidden markov model based algorithms for these purposes)
- Principles and applications of homology, fold recognition, and ab initio based algorithms for teritairy structure prediction of proteins, application of protein tertiary structure prediction towards problems of protein folding and design
- Learn and apply the principles of molecular modelling and energy minimization for small molecule –protein and protein-protein binding ; learn the principles and methodologies of computer aided drug design.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*